DOCUMENT RESUME

ED 468 254 SP 041 014

AUTHOR	Gallagher, H. Alix
TITLE	The Relationship between Measures of Teacher Quality and Student Achievement: The Case of Vaughn Elementary.
SPONS AGENCY	National Inst. on Educational Governance, Finance, Policymaking, and Management (ED/OERI), Washington, DC.; Wisconsin Center for Education Research, Madison.; Consortium for Policy Research in Education.
PUB DATE	2002-04-00
NOTE	32p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
CONTRACT	OERI-R308A60003
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; Disadvantaged Youth; Elementary Education; High Risk Students; Knowledge Base for Teaching; *Performance Based Assessment; Reading Skills; Teacher Competencies; *Teacher Evaluation; Teacher Knowledge; Teaching Skills; Urban Schools

ABSTRACT

This paper reports on a study of the relationship between teacher evaluation scores in a school implementing knowledge- and skills-based pay and classroom student achievement. The study occurred in a California charter elementary school that was 100 percent Title I, 100 percent free/reduced lunch, and had predominantly limited English speaking students. The school had historically low achievement, and for 4 years, it had been implementing a performance evaluation and pay plan under which teachers were evaluated, rated, and paid accordingly. For the study, data were collected on 34 teachers and all of their students for whom 2 years of achievement data were available. Researchers estimated classroom effects, analyzed their relationship to teacher evaluation scores, and examined teacher evaluation scores as level 2 explanatory variables in hierarchical linear models of student achievement. Results indicated that there was a clear difference in the strength of the relationship between teacher evaluation scores and classroom achievement in reading compared with mathematics or language arts. An appendix presents descriptive statistics. (Contains 37 bibliographic references.) (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

CONSORTIUM FOR POLICY RESEARCH IN EDUCATION
University of Pennsylvania • Harvard University • Stanford University
University of Michigan • University of Wisconsin-Madison

ED 468 254

The Relationship Between Measures of Teacher Quality and Student Achievement: The Case of Vaughn Elementary

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

H. Alix Gallagher

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

H. Alix Gallagher

University of Wisconsin
1025 W. Johnson St.
Madison, WI 53706
608.265.3523
hagallagher@students.wisc.edu

April 2002

Prepared for the 2002 annual meeting of the
American Educational Research Association
New Orleans, Louisiana

This paper was prepared for the Consortium for Policy Research in Education, Wisconsin Center for Education Research, University of Wisconsin-Madison for presentation at the American Educational Research Association annual conference held April 1-5 in New Orleans, Louisiana. The research reported in this paper was supported by a grant from the U.S. Department of Education, Office of Educational Research and Improvement, National Institute on Educational Governance, Finance, Policymaking and Management, to the Consortium for Policy Research in Education (CPRE) and the Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison (Grant No. OERI-R308A60003). The opinions expressed are those of the authors and do not necessarily reflect the view of the National Institute on Educational Governance, Finance, Policymaking and Management, Office of Educational Research and Improvement, U.S. Department of Education, the institutional partners of CPRE, or the Wisconsin Center for Education Research.

SP041014

UNIVERSITY OF WISCONSIN MADISON

Wisconsin Center for Education Research, University of Wisconsin-Madison
1025 West Johnson Street, Room 659, Madison, WI 53706-1796 ☐ Phone 608.263.4260 ☐ Fax 608.263.6448

BEST COPY AVAILABLE

The main goal of standards-based reform is to improve student outcomes by focusing on student achievement. Many researchers and policy-makers have noted that to improve student learning, teachers will need to increase their skills. One strategy advanced for motivating teachers to acquire the capacity necessary to achieve the goals of standards-based reform is knowledge- and skills-based pay, which attaches financial rewards to teacher demonstration of specific competencies (Odden & Kelley, 2002). The effectiveness of a system obviously rests on implementation and also whether or not the teacher evaluation system rewards teacher knowledge and skills that contribute to student achievement. This paper reports on a study of the relationship between teacher evaluation scores in a school implementing knowledge- and skills-based pay, and classroom student achievement. The study estimates 'classroom effects,' analyzes their relationship to teacher evaluation scores, and examines teacher evaluation scores as level 2 explanatory variables in hierarchical linear models of student achievement.

Theoretical Background

The question of how or whether teachers impact student achievement is often framed against Coleman's *Equality of Educational Opportunity*, frequently referred to as the Coleman Report, released in 1966 (Coleman, 1990). The Coleman report used aggregated measures of school inputs in terms of facilities, teacher characteristics (specifically the average educational level of the teachers' families, average years of experience, on average whether teachers were local to the area, the teachers' average level of education, the teachers' average scores on a self-administered vocabulary test, the teachers' preference for teaching middle-class, white-collar students, the proportion of teachers in the school who were white), and student population characteristics to investigate the effect of schools on students' educational achievement. The report showed that schools' average student characteristics, such as poverty and attitudes towards schooling, often had a greater impact on student achievement than teachers and schools. In a 1990 book, *Equality of Achievement in Education*, Coleman discussed the findings of his earlier work. He summarized them by saying, "The principal result, based on a variety of analyses, is as follows: *Attributes of other students account for far more variation in the achievement of minority*

BEST COPY AVAILABLE

group children than do any attributes of school facilities and slightly more than do attributes of staff." (Coleman, 1990, p. 86, italics in the original). Furthermore, he noted:

> School to school variations in achievement, from whatever source (community differences, variations in the average home background of the student body, or variations in school factors), are much smaller than individual variations within the school, at all grade levels, for all racial and ethnic groups. ....[T]he factors that, under all conditions, accounted for more variance than any others were the characteristics of the student's peers; those that accounted for the next highest amount of variance were teachers' characteristics; and finally, other school characteristics, including per pupil expenditure on instruction in the system, accounted for very little variance at all. The total variance accounted for by these three sets of school factors was not large—in fact, an analysis of variance showed that only about 10 percent of the variance in achievement lay between schools. (Coleman, 1990, pp. 77)

This report led some to believe that schools and teachers did not have a significant impact on student outcomes (Porter & Brophy, 1988). Coleman's (1990) findings can be somewhat deceptive since his study relied on data aggregated to the school level. He found that average teacher characteristics at a school had a small impact on a school's mean achievement; this should not be interpreted to mean that individual teachers or structures within schools do not have a large impact on student achievement. By using only aggregate data, Coleman eliminated variation within schools and, therefore, Coleman's study design did not enable him to look at individual teachers' effects. Additionally, as Dreeben (2000) points out, Coleman did not investigate the mechanisms by which any properties of schools (or classrooms) influence *individual* student achievement.

A significant body of research now stands in support of Coleman's findings that external characteristics (such as student socio-economic status [SES] and parental educational attainment) impact student achievement in significant ways (Meyer, 1996; Porter & Smithson, 2000; Webster, Mendro, Orsak, Weerasinghe, & Bembry, 1997); but when those differences are controlled for, teachers are the most important determinants of student achievement (Brophy, 1986; Darling-Hammond, 2000; Haycock, 1998; Webster, Mendro, Orsak, & Weerasinghe, 1998; Wright, Horn, & Sanders, 1997). The important role of teaching in facilitating high levels of student achievement on basic skills tasks was demonstrated conclusively by the effective school research (Brophy, 1986; Cohen, 1983). More recently a large body of knowledge has been developed that shows the

important role of high quality teaching in achieving the more complex learning goals of standards-based reform (Bransford, Brown, & Cocking, 1999; Darling-Hammond, 2000; Darling-Hammond & Ball, 1998; Rowan, Chiang, & Miller, 1996).

Based on the understanding that teacher competency is important for student outcomes, knowledge- and skills-based pay seeks to provide an extrinsic incentive for teachers to acquire important skills. Unlike merit pay systems, which were common in earlier attempts to reward outstanding teachers, knowledge- and skills-based pay systems do not foster competition amongst teachers because they pay all teachers who demonstrate the desired skills and competencies (Odden & Kelley, 2002). Knowledge- and skills-based pay can help to improve student achievement if several conditions are met: 1) criteria for teacher knowledge and skills are developed to focus teachers on acquiring proficiency in key areas; 2) teachers are evaluated on those criteria; 3) teachers are motivated by the system to acquire desired skills. Additionally a knowledge- and skills-based pay system that seeks to improve outcomes rests on the key assumption that teachers who receive higher teacher evaluation scores produce greater growth in student achievement than teachers who receive lower evaluation scores. This study tests the validity of that assumption in the case of Vaughn Elementary School, a pre-K-5 urban elementary school in Los Angeles.

It is important to note that the predictive validity of measures of teacher quality has proved challenging to the field. A long research trajectory has used teacher production functions to examine differences in student outcomes based on many teacher characteristics including: years of experience, certification, advanced degrees, verbal ability, and many others for example (Darling-Hammond, 2000; Greenwald, Hedges, & Laine, 1996; Hanushek, 1971). Although this body of research is too large to fully review here, it is fair to say results have been mixed.

Results of research that have tried to tie teacher evaluation to student achievement have shown that principals' evaluations of teachers, the most common form of teacher evaluation, typically have little to no correlation with student achievement (Medley & Coker, 1987; Peterson, 2000). As Medley notes discussing research in this area, "...additional studies of this problem were

published, all of which reached the same conclusions: that the correlations between the average principal's ratings of teacher performance and direct measures of teacher effectiveness were near zero." (Medley & Coker, 1987, p. 242) This has led to calls for improvements in principals' evaluation of teachers (Nelson & Sassi, 2000; Peterson, 2000).

Some of the more recent improvements in teacher evaluation can be found in the area of teacher certification, where several promising evaluations are being developed for teachers at different career stages. One of the most high profile examples of this is the certification process created by the National Board of Professional Teaching Standards [NBPTS]. Research on NBPTS certification has demonstrated that it meets criteria of reliability, consistency and certain types of validity (Porter, Youngs, & Odden, 2001). However, a study by Bond et al. (Bond, Smith, Baker, & Hattie, 2000) showed that while students of NBPTS certified teachers outperformed students whose teachers were not NBPTS certified on curriculum embedded assessments, there was no significant difference in their performance on external measures. Clearly for systems to attach significant monetary incentives to teacher evaluation, this type of validity needs to be established. The examples of principal evaluation and NBPTS show how challenging this task is in education.

It is important to note that convergent validity between supervisory evaluations and results-oriented measures of performance has been shown to be relatively low across fields. Heneman's (1986) meta-analysis of private sector performance appraisal found that the correlation between supervisory evaluations and employee performance outcomes average .27. While this at first seems low, factors such as differences between the aspects of the job on which the employee is evaluated and the result outcome, in addition to evaluator biases, would potentially decrease the correlation. Using the example of teaching, most people would argue that teaching requires a variety of proficiencies that can justifiably contribute to teacher evaluations, yet which may only indirectly influence student performance on a given assessment. Therefore, one might not expect a higher correlation.

Site Selection

Since this paper seeks to examine the validity of an evaluation system embedded in knowledge- and skills-based pay, the first step was to identify a good case. Vaughn Elementary was chosen because it had been involved in implementing and annually refining a fairly sophisticated knowledge- and skills-based pay system for over three years.

Vaughn Elementary is a charter school in the Los Angeles Unified School District serving approximately 1200 students. The school is 100% Title I, 100% free/reduced lunch, and 85% of its student body is classified English language learners. Prior to receiving a charter in 1993, Vaughn had very low student achievement, with many students scoring in the lowest 10th percentile on norm-referenced tests. In its charter, Vaughn listed improving student performance as a critical goal and measure of success. Student performance has improved substantially since obtaining charter status and the school has been recognized nationally as a National Blue Ribbon School and has qualified for a school performance bonus under the California Academic Performance Index (a reward component of the state accountability system based on comparing student test scores at demographically similar schools) during the 1999-2000 school year (Kellor, Milanowski, Odden, & Gallagher, 2001; Milanowski & Gallagher, 2001; WestEd, 1998).

For the past four years, Vaughn has been implementing a performance evaluation and pay plan. Under this system teachers are evaluated during three, week-long windows throughout the school year across up to ten domains: lesson planning, classroom management, literacy, mathematics, language development, special education inclusion, social studies, science, art, and technology. In each domain they are evaluated on several teaching standards by an administrator, a trained peer and themselves. Teachers are rated on a scale of 1 (unsatisfactory) to 4 (exemplary) on each standard. The ratings are averaged to form a rating in each domain, which are then averaged to an overall rating. Additionally if teachers have a bilingual credential [BCLAD] they receive a further stipend depending on their students' English language proficiency. Like many other urban schools in California, Vaughn has a substantial number of teachers working on an emergency teaching

BEST COPY AVAILABLE

credential. Vaughn evaluates these teachers only in the first five areas, which they consider to be most critical. Fully certified teachers, who achieve an overall average rating of 3.0 are evaluated in all ten areas. (The complete rubrics can be found on the CPRE website).

Methods

This paper reports on the first part of a two-part study. The current study used correlations between residuals for classroom performance and teacher evaluation scores and hierarchical linear modeling to examine the relationship between teacher evaluation scores and value-added measures of student achievement at Vaughn. The second part of the study, not reported here, used interviews and document analysis to explore the evaluation system in more detail.

The sample for this study was thirty-four 2nd-5th grade teachers and all of their students for whom two years of student achievement data were available. The measure of teacher performance used was teachers' average score in each domain across the three observation windows. Although Vaughn did not keep evaluation data from the 2000-2001 school year in a form that enabled the researcher to calculate the inter-rater reliability, such an analysis was conducted for the fall 2001 semester. Additionally, data was collected on other teacher characteristics, including teachers' years of experience and certification status, which were used in additional analyses not reported on here. There was not enough variation in the sample to allow modeling of the impact of teachers holding advanced degrees. Appendix 1 contains descriptive statistics for all teacher variables. Table 1 shows the reliability, mean, standard deviation and range for the teacher evaluation scores.

Table 1: Descriptive Statistics for Teacher Evaluation Scores

Variable	Alpha coefficient for Reliability	Mean	Standard Deviation	Minimum	Maximum
Average Literacy TES	.81	3.24	0.39	2.30	3.85
Average Math TES	.83	3.13	0.42	2.25	4.00
Average Language Development TES	.83	3.24	0.42	2.30	3.95

As is apparent from Table 1, all teacher evaluation scores had an alpha coefficient greater than .81. This is relatively high compared to findings from a meta-analysis, which found an average internal consistency of .60 for supervisory ratings (Heneman, 1986). Most teachers scored in a range from

approximately 2.7 to about 3.7 in the various categories. Given the potential range of scores from 1.0 to 4.0 this is a relatively narrow distribution.

The Stanford-9 [SAT-9] was used as the measure of student achievement because it is the indicator of student performance used by the state accountability policy, and is thus likely to be more aligned with the curriculum than a researcher-developed instrument. This importance of alignment has been demonstrated conclusively by Porter and Smithson (2000), who show that the alignment between taught and tested curriculum, both in terms of content and cognitive demand, is a highly significant predictor of student performance. Although there have been some criticisms about the degree of alignment between the SAT-9 and the California Standards, the California version of the SAT-9 has been augmented to better reflect standards. Furthermore, the accountability system is designed to encourage alignment with the SAT-9; in as much as teachers pay attention to this policy pressure, the SAT-9 is the best instrument for measuring student performance (Herman, Brown, & Baker, 2000). However, no single measure should be seen as the sole criterion for judging performance; this is especially true given the additional assessment issues raised by testing Vaughn's large population of English language learners in English (August & Hakuta, 1998; Heubert & Hauser, 1999). All conclusions in this paper need to be understood within the limitations of the measurement instruments.

As was suggested by prior research, the models controlled for student characteristics. Appendix 1 contains a list of all control variables and their measures. The study used a pre-test/post-test model (Spring 2000/Spring 2001 testing) to control for students' prior achievement. Value-added models such as these are useful for this type of analysis because they isolate growth in achievement during a specific time period and allow the researcher to control for student characteristics that are related to student outcomes (Meyer, 1996, 1997). Although this study did not employ individual student gain scores in the models, Table 2 below reports average student gain scores for each subject to help the reader to understand the coefficients presented in the models that

follow. Gains are reported for each of the subject tests, and a composite measure of achievement across all three tested areas.

Table 2: Average Student Gain for SAT-9 from Spring 2000 to Spring 2001

Subject	Mean	Standard Deviation	Minimum	Maximum
Reading	32.776	24.574	-42.00	127.00
Math	34.732	24.955	-33.00	140.00
Language Arts	22.430	26.543	-44.00	137.00
Composite	29.980	19.659	-27.00	92.67

These results show that an average gain across all three subjects for a student in one year is slightly less than thirty points, with gains in math and reading higher than gains in language arts. Thirty points will thus be used an as approximate benchmark of an average year's progress at Vaughn.[1] The extreme minimum and maximum scores, also suggest that the sample has some of the volatility of scores mentioned in some of the literature (Rogosa, 1999).

Hierarchical linear modeling was selected for the core analyses because it takes advantage of the nested nature of the data set where students are grouped in classes for instruction (Bryk & Raudenbush, 1988). Bayes' Estimates of the residuals for each classroom's intercept were used to estimate 'classroom effects.' Unlike previous studies that use the term 'teacher effects,' this study uses 'classroom effects' because it more accurately describes what is being estimated: the group level residual for a given class after controlling for certain individual and group characteristics.

This study addressed two main questions:

1. What is the 'classroom effect' in each subject and how does the Vaughn measure of teacher quality correlate with these student outcomes?
2. Were Vaughn's teacher evaluation scores significant predictors of variation in student achievement?

[1] Harcourt Brace, the publisher of the Stanford-9, would not release norms and reliability information about the California form of the Stanford-9, and so the author could not compare growth of students at Vaughn to a sample that was representative of either the nation or the State of California.

BEST COPY AVAILABLE

Estimates were derived for reading, math, language arts and overall. Each analysis began with a random intercept hierarchical linear model, which included all theoretically relevant level 1 student characteristics control variables. This model can be expressed generally as:

Level 1: $\text{Post-test}_{ij} = \beta_{0j} + \theta\text{Pre-test}_{ij} + \alpha\text{studchar}_{ij} + \varepsilon_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + U_{0j}$
$\beta_{kj} = \gamma_{k0}$

In this model, a Bayes' estimate of U_{0j} was used to estimate the classroom effect. Bayes estimates of U_{0j} were then correlated with teacher evaluation scores to determine the relationship between classroom effects and teacher evaluation scores. The random intercept model is useful because estimates of τ_{00} and σ^2 can be used to calculate the intraclass correlation coefficient $\rho = \tau_{00}/(\tau_{00}+\sigma^2)$, which describes the proportion of variance at the group (classroom) and individual (student) levels, an important concept in studies of multilevel effects and outcomes.

Next, the teacher evaluation scores were inserted into the model as a level 2 predictor. This can be expressed generally as:

Level 1: $\text{Post-test}_{ij} = \beta_{0j} + \theta\text{Pre-test}_{ij} + \alpha\text{studchar}_{ij} + \varepsilon_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{TES})_j + U_{0j}$
$\beta_{kj} = \gamma_{k0}$

The estimates of γ_{01} were examined for both statistical and practical significance. As can be seen from the equations, this study used random intercept models because there were no statistical differences in slopes across classes. The next section presents the results of the correlation analyses for each subject separately, followed by a comparison of how the different subjects' teacher evaluation scores performed as predictors of class level variation in student achievement.

Results

In working with HLM, it is critical to have accurately specified level 1 of the model, otherwise all level 2 estimates, which are the focus of this study, could be biased. As a result, all subject-area analyses began with all level one variables entered into the model. After those

preliminary analyses, gender was dropped from all analyses because it was both statistically insignificant and lacked theoretical grounding. Preliminary analyses also used dummy variables for students' 2001 grade level to examine whether there were differences across grades. The dummy variables were statistically insignificant, theoretically undesirable and had small coefficients, so grade level was also dropped from the analysis. Level 2 variables representing aggregate student characteristics for average pretest and average English language proficiency were also explored to ensure that the level 1 controls for student characteristics adequately accounted for differences in these variables. They were insignificant and so they were dropped from the model. Finally, to handle missing data in Level 1 reading and math scores, conditional mean imputation was used. For missing data on students' English language proficiency, unconditional class mean imputation was used since students were grouped by language proficiency. A dummy variable was created in all cases to denote imputed data (Little & Rubin, 1987).

Both reading and math scores from the Spring 2000 testing were included in the model to increase the reliability of the understanding of students' prior performance. The resulting model was:

Level 1: $\text{ScaledReading01} = \beta_0 + \beta_1(\text{Attendance})_{ij} + \beta_2(\text{ImputeAttendance})_{ij} + \beta_3(\text{Retain})_{ij} + \beta_4(\text{SpecialEducation})_{ij} + \beta_5(\text{ScaledReading00})_{ij} + \beta_6(\text{ScaledMath00})_{ij} + \beta_7(\text{ImputedReading})_{ij} + \beta_8(\text{ImputedELD})_{ij} + \beta_9(\text{EarlyELD})_{ij} + \beta_{10}(\text{EnglishProficient})_{ij} + \beta_{11}(\text{ParentsNoHS})_{ij} + \beta_{12}(\text{MissingParentEd})_{ij} + R_{ij}$

Level 2: $\beta_0 = \gamma_{00} + U_{0j}$

β_1 through β_{12} are fixed.

The descriptive statistics for all variables in the reading analysis are in Appendix 1. The results of the random intercept model used to generate the empirical Bayes' estimates is presented in Table 3 below. The table is accompanied by an analysis of the different coefficients. This detailed analysis is provided only for reading, with this analysis serving as a model of analysis for the other subjects.

Not surprisingly, individual pre-test scores are strong predictors of post-test performance in reading. For every point higher the student scored on the reading pre-test, they scored .5904 points higher on post test; for every point higher the student scored on the math pre-test, they scored .1518 points higher on the reading post test, controlling for all other factors.

Table 3: Results for the random intercept model for reading

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
Intercept G00	181.9305	15.9190	11.429	0.000 ***
Attendance (grand centered) G10	-0.0258	0.0900	-0.287	0.774
Impute Attendance G20	6.0684	8.5880	0.707	0.480
Retain G30	-8.3977	3.3842	-2.481	0.013 **
Special Education G40	-23.7429	4.0970	-5.791	0.000 ***
Scaled Reading pre-test G50	0.5904	0.0620	18.0425	0.000 ***
Scaled Math pre-test G60	0.1518	0.0323	4.700	0.000 ***
Impute Reading G70	-7.1873	3.4624	-2.076	0.038 **
Impute ELD G80	-3.3232	2.2200	-1.497	0.134
Beginning ELD G90	-8.7192	3.7268	-2.340	0.019 **
English Proficient G100	3.2783	2.6126	1.255	0.210
Parent No HS G110	-1.109	1.9880	-0.558	0.576
Missing parent education G120	-4.3422	2.5035	-1.734	0.082 *
Reliability		**Variance Components**		
Reliability of $\beta_0 = 0.747$		$\tau_{00} = 59.2020$, chi-square p-value 0.000		
Reporting		$\sigma^2 = 339.4203$		
*** = p < .001, ** = p < .01, * = p < .1		p = 0.14852		
Scale: a one point gain in reading on the Spring 2001 test.				
Note-in all tables G refers to γ				

Additionally, students who were in special education grew less than their peers in regular education. Also unsurprising was that for students who are at an beginning level of English Language Development, scores were lower than peers with more advanced English skills. The difference between being fully English proficient, as opposed to intermediate or advanced intermediate, was not statistically significant. Being retained also had a large and significant coefficient. This is especially surprising since the results show that for students experiencing a grade for the *second* time, scaled scores grew an average of over eight points less than their peers experiencing the grade for the *first* time.

Students with imputed reading scores had slightly less growth than their peers. Two potential explanations for this are that the imputations may be a slightly biased estimate of performance, or that students who are less successful are less likely to complete all sections of the reading test. To test whether imputed reading scores impacted the outcomes of the analyses, the researcher generated empirical Bayes' estimates of classroom effects of a model that did not include

BEST COPY AVAILABLE

the 43 students who had imputed reading scores. The correlation between the Empirical Bayes' classroom effects with and without imputed scores was .994, so the students were left in the model.

While parental educational attainment does not seem significant, students for whom no data was available showed less growth than their peers. Since these cases were a combination of students whose parents declined to respond and students who had left Vaughn when this data was collected, it could also reflect lower (though not statistically significant) growth in last year's fifth graders (compared to other grades) or a creaming effect of the school. In either case, the finding is only marginally statistically significant, and should be interpreted with more caution than the other findings. Finally, the researcher decided to leave student attendance and the dummy variable for imputed attendance in the model even though neither were significant, because attendance is typically used in such studies and, theoretically, there is a strong relationship between attendance and student achievement.

A reliability of 0.747 for β_0 suggests that this model generates a relatively good estimate of the mean. The chi-square test for random effects of U_0 was significant, with a p-value of 0.000. This indicates that there is statistically different variation in achievement between classes. The intraclass correlation coefficient of .1485 shows that approximately 15% of the variation in student achievement is between classes. This is within the range of earlier research on classroom effects (Rowan, 2001).

The next step of the analysis was to examine the correlation between Bayes' estimates of classroom effects in reading and teacher evaluation scores. Prior to generating the Bayes' estimates, several level 2 variables were tested for significance: class average reading pre-test and class average ELD. Both were found to be statistically insignificant, implying that level 1 of the model was a sufficient control for variation in students' prior performance and English Language Development so that teachers with low performing or low ELD students were not penalized in the model. As a result, the Bayes' estimates were generated from a model with no level 2 predictors. Table 4 reports the Pearson correlations between the Bayes' estimates and teachers average literacy score, their

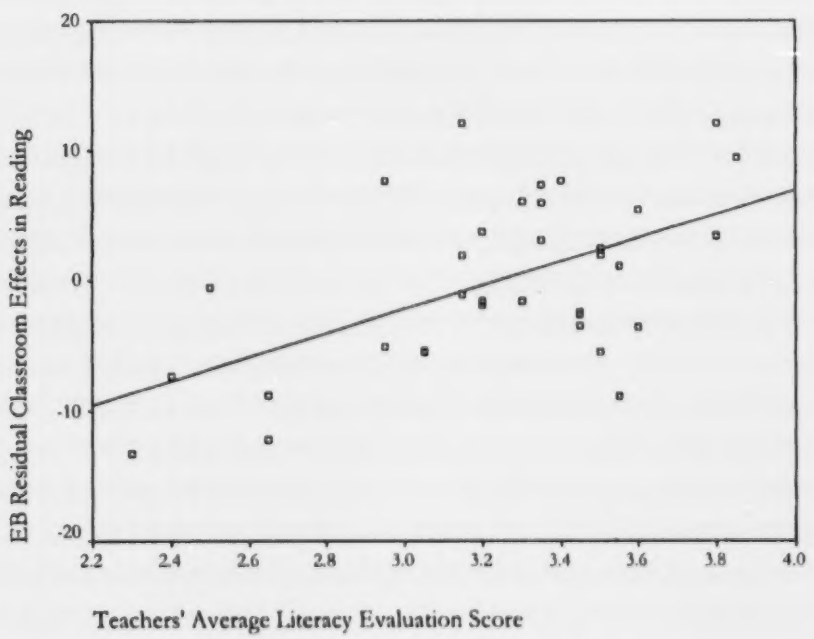
average score across the 5 core domains (lesson planning, classroom management, literacy, mathematics, and classroom development), and their overall average score (all domains on which teachers were rated).

Table 4: Correlation of Empirical Bayes' estimates of classroom effects in reading with indicators of teacher quality

	Average Literacy	Average 5 core domains	Overall average
Pearson Correlation	.545 **	.483 **	.549 **
** Correlation is significant at the 0.01 level (2-tailed).			

These are in the range of correlations that would be expected from a high quality performance evaluation (Heneman, 1986). Figure 1 below represents this correlation graphically for the teachers' average literacy rating, the variable which is of the most theoretical interest. The x-axis represents teachers' average evaluation score in literacy over the course of the 2000-2001 school year. The y-axis, represents Empirical Bayes' residuals of classroom effects. The superimposed fit line shows that the correlation is positive, yet the spread of individual observations around the line shows that the fit, while inexact, is still clearly linear.

Figure 1: The correlation between Literacy TES and Student Growth in Reading



These results show that there is a moderately high correlation between student performance and teacher evaluation scores overall, in literacy, and across the five key domains.

The study next turned to entering teacher characteristics into level 2 of the model to provide more information about teacher evaluation scores as predictors of growth in student achievement. Such models are useful because they also address the question of most practical significance, 'What is the affect on student achievement for students who had a top performing, as opposed to a low performing, teacher?' This question was examined using the following model, which is identical to the random intercept model used earlier with the exception of the addition of the teachers' literacy evaluation scores at level 2:

Level 1: $Y = \beta_0 + \beta_1(\text{Attendance})_{ij} + \beta_2(\text{ImputeAttendance})_{ij} + \beta_3(\text{Retain})_{ij} + \beta_4(\text{SpecialEducation})_{ij} + \beta_5(\text{ScaledReading00})_{ij} + \beta_6(\text{ScaledMath00})_{ij} + \beta_7(\text{ImputedReading})_{ij} + \beta_8(\text{ImputedELD})_{ij} + \beta_9(\text{EarlyELD})_{ij} + \beta_{10}(\text{EnglishProficient})_{ij} + \beta_{11}(\text{ParentsNoHS})_{ij} + \beta_{12}(\text{MissingParentEd})_{ij} + R_{ij}$

Level 2: $\beta_0 = \gamma_{00} + (\text{AvgLiteracyTES}) + U_{0j}$
β_1 through β_{12} are fixed.

Table 5 reports the results of that analysis.

Table 5: Literacy TES as a predictor of student performance in reading

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
Intercept G0	188.4932	15.3613	12.271	0.000
Average Literacy TES (grand centered) G01	12.9265	3.5397	3.652	0.001 ***
Attendance (grand centered) G10	-0.0380	0.0896	-0.425	0.671
Impute Attendance G20	6.0847	8.5564	0.711	0.477
Retain G30	-8.0534	3.3733	-2.387	0.017 **
Special Education G40	-23.7786	4.0794	-5.829	0.000 ***
Scaled Reading pre-test G50	0.5818	0.0315	18.456	0.000 ***
Scaled Math pre-test G60	0.1488	0.0320	4.649	0.000 ***
Impute Reading G70	-6.4844	3.4092	-1.902	0.057 *
Impute ELD G80	-2.9800	2.1722	-1.372	0.170 **
Beginning ELD G90	-8.5392	3.6386	-2.347	0.019 *
English Proficient G100	2.8518	2.5543	1.116	0.265
Parent No HS G110	-1.2037	1.9778	-0.609	0.542
Missing parent education G120	-4.6212	2.4235	-1.907	0.056 *
Reliability		**Variance Components**		
Reliability of $\beta_0 = .655$		$\tau_{00} = 37.9265$		
Reporting *** = p < .001, ** = p < .01, * = p < .1		$\sigma^2 = 338.9549$		
Scale: a one point gain in reading on the Spring 2001 test. **Note**-in all tables G refers to γ		**Proportion of between class variation explained by reading TES** = .3594		

Since level 1 coefficients were interpreted for the earlier model, a full discussion of level 1 coefficients will not be provided here or in any subsequent section. The most important finding for this analysis is that teachers' average evaluation scores in literacy are a highly statistically significant predictor of student performance. For every point increase in teacher evaluation scores, student performance increases almost thirteen points.

Perhaps a more useful way to look at this, given the relatively small variation in teacher evaluation scores in reading, is to note that for every standard deviation of improvement in literacy teacher evaluation score (0.39 = one standard deviation), student performance improves 5.041 points. The range in teachers' evaluation scores in literacy is 1.55 points, so the difference for the average student in the top performing class and the bottom performing classroom, after controlling for all other factors, would be predicted to be 20.035 points a year. This represents slightly less than two-thirds of the average growth in reading for Vaughn students from 2000 to 2001, and is thus a practically significant result.

Another way to look at the effectiveness of the literacy teacher evaluation score is to see how much of the between class variation in student achievement that was represented in the random intercept model is explained by the teacher evaluation score. The formula for the proportion of variance explained is:

Prop of Var expl. L2 = [$\hat{t}_{00}$(random int.) - $\hat{t}_{00}$ (Level 2 pred. Model)] / $\hat{t}_{00}$(random int)

Calculations here are: (59.20202-37.92654)/59.20202 = .3594 or almost 36% of true between class variance in reading achievement. This study now turns to a similar discussion of math.

Math

As with the reading analysis, prior to determining the final model, the researcher examined assumptions of normality and checked to make sure that indicators for gender and grade level could be dropped. The resulting sample of 584 students was identical to the sample for the reading analysis. Additionally, as with reading, the researcher tested level 2 variables for class average ELD and class average pretest to make sure that the coefficients were small and insignificant. Both

reading and math scores from the Spring 2000 testing were included in the model to increase the reliability of the understanding of students' prior performance. The resulting model was:

Level 1: $ScaledMath01 = \beta_0 + \beta_1(Attendance)_{ij} + \beta_2(ImputeAttendance)_{ij} + \beta_3(Retain)_{ij} + \beta_4(SpecialEducation)_{ij} + \beta_5(ScaledReading00)_{ij} + \beta_6(ScaledMath00)_{ij} + \beta_7(ImputedReading)_{ij} + \beta_8(ImputedELD)_{ij} + \beta_9(EarlyELD)_{ij} + \beta_{10}(EnglishProficient)_{ij} + \beta_{11}(ParentsNoHS)_{ij} + \beta_{12}(MissingParentEd)_{ij} + R_{ij}$

Level 2: $\beta_0 = \gamma_{00} + U_{0j}$

β_1 through β_{12} are fixed.

The descriptive statistics for all variables in the math analysis are in Appendix 1. The results of the random intercept model used to generate the empirical Bayes' estimates is presented in Table 6.

Table 6: Results for the random intercept model for math

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
Intercept G00	161.1311	18.3797	8.767	0.000 ***
Attendance (grand centered) G10	0.2126	0.0986	2.156	0.031 **
Impute Attendance G20	17.5968	9.4173	1.869	0.061 *
Retain G30	-1.4233	3.7089	-0.384	0.701
Special Education G40	-17.1520	4.5030	-3.809	0.000 ***
Scaled Reading pre-test G50	0.1971	0.0360	5.482	0.000 ***
Scaled Math pre-test G60	0.5944	0.0356	16.703	0.000 ***
Impute Reading G70	-2.4336	3.8466	-0.633	0.527
Impute ELD G80	-3.3256	2.4804	-1.341	0.180
Beginning ELD G90	-0.8078	4.1714	-0.194	0.847
English Proficient G100	-0.8165	2.9198	-0.28	0.782
Parent No HS G110	-0.6468	2.1828	-0.296	0.767
Missing parent education G120	-3.6633	2.8335	-1.293	0.196
Reliability		**Variance Components**		
Reliability of $\beta_0 = 0.832$		τ_{00}= 118.41321, chi-square p-value 0.000		
Reporting *** = p < .001, ** = p < .01, * = p < .1		σ^2 = 405.71662		
Scale: a one point gain in reading on the Spring 2001 test. **Note**-in all tables G refers to γ		ρ = .22592		

Since the coefficient for imputed attendance was large and significant, the researcher created a separate file with those cases deleted and generated empirical Bayes' residuals for the intercept. Since these correlated at a .999 level with the residuals generated with the larger sample, the five students with imputed attendance were retained in the sample for the rest of the analysis.

Though the coefficient for attendance is relatively small, it is significant, which confirms the prudence of leaving it in all the models to avoid generating biased estimates. For every day more

BEST COPY AVAILABLE

than the average that students were in school, their math scores went up approximately .2 points, which is not practically significant for most students. However, for the fifteen students who missed twenty or more days of school, this could be practically significant because it reflects an average of one point lower performance for every five days missed. No other level 1 coefficients in this analysis are worth a lengthy discussion.

The reliability coefficient for β_0 was 0.832, reflecting a good fit. Like in the reading analysis, the chi-square test for random effects of U_0 was significant, with a p-value of 0.000. This indicates that there is statistically different growth in achievement between classes. The intraclass correlation ρ was .22592, meaning that almost 23% of the variation in student performance was between classes, with 77% of the variation between individuals.

Next the analysis examined the correlations between the Empirical Bayes' estimates of classroom intercepts and the teachers' average math evaluation score, average score across the five core domains, and overall average score. Table 7 shows the results.

Table 7: Correlation of Empirical Bayes' estimates of classroom effects in mathematics with indicators of teacher quality

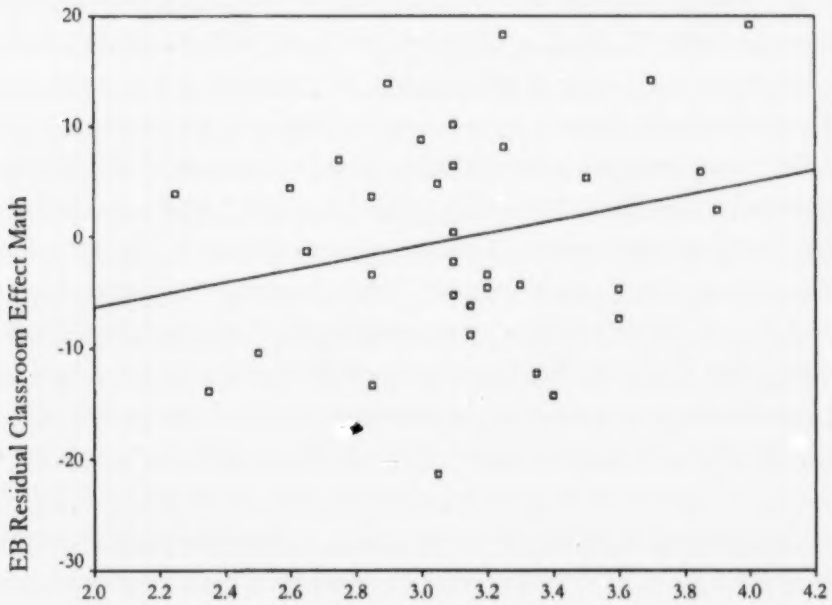
	Average Math	Average 5 core domains	Overall average
Pearson Correlation	.239	.204	.239

In contrast to the reading analysis, no correlations were statistically significant. This can be seen easily by looking at the scatterplot in Figure 2 below, where teachers' average mathematics evaluation scores are plotted on the x-axis, and the Empirical Bayes' residual for their classrooms' math performance is plotted on the y-axis.

As the scatterplot with the superimposed overall fit line shows, there is a slight positive correlation between teacher evaluation scores and student achievement in math. However, a comparison between Figures 1 (reading) and 2 (math) show that the spread of scores around the fit line is much broader in math than in reading. In fact, the teacher whose students had the worst

performance, with an Empirical Bayes' residual of –21.40, received a math evaluation score of 3.05, which is just below the mean of 3.13.

Figure 2: The correlation between Math TES and Student Growth in Math



To further investigate mathematics teacher evaluation scores as a predictor of student performance in math, they were entered into the model as a level 2 predictor in the following equation:

Level 1: $(ScaledMath01) = \beta_0 + \beta_1(Attendance)_{ij} + \beta_2(ImputeAttendance)_{ij} + \beta_3(Retain)_{ij} + \beta_4(SpecialEducation)_{ij} + \beta_5(ScaledReading00)_{ij} + \beta_6(ScaledMath00)_{ij} + \beta_7(ImputedReading)_{ij} + \beta_8(ImputedELD)_{ij} + \beta_9(EarlyELD)_{ij} + \beta_{10}(EnglishProficient)_{ij} + \beta_{11}(ParentsNoHS)_{ij} + \beta_{12}(MissingParentEd)_{ij} + R_{ij}$

Level 2: $\beta_0 = \gamma_{00} + (AvgMathTES) + U_{0j}$

β_1 through β_{12} are fixed.

Table 8 reports those results.

BEST COPY AVAILABLE

Table 8: Math TES as a predictor of student math performance

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
Intercept G00	164.0395	18.4257	8.903	0.000 ***
Average Math TES (grand centered)	6.9685	4.9854	1.398	0.172
Attendance (grand centered) G10	0.2104	0.0986	2.134	0.033 *
Impute Attendance G20	17.6000	9.4162	1.869	0.061 *
Retain G30	-1.4198	3.7086	-0.383	0.701
Special Education G40	-17.1842	4.5020	-3.817	0.000 ***
Scaled Reading pre-test G50	0.1938	0.0360	5.387	0.000 ***
Scaled Math pre-test G60	0.5926	0.0356	16.646	0.000 ***
Impute Reading G70	-2.1296	3.84755	-0.553	0.579
Impute ELD G80	-3.2185	2.4783	-1.299	0.194
Beginning ELD G90	-0.8258	4.1652	-0.198	0.843
English Proficient G100	-0.7849	2.9160	-0.269	0.788
Parent No HS G110	-0.7344	2.1834	-0.336	0.736
Missing parent education G120	-3.5783	2.8273	-1.266	0.206
Reliability		**Variance Components**		
Reliability of $\beta_0 = .826$		$\tau_{00} = 113.91646$		
Reporting *** = p < .001, ** = p < .01, * = p < .1		$\sigma^2 = 405.7859$		
Scale: a one point gain in reading on the Spring 2001 test. **Note**-in all tables G refers to γ		**Proportion of between class variation explained by math TES** = .0380		

The average math teacher evaluation score is not a statistically significant predictor of student performance in math. Since the sample at level 2 is relatively small, it is also worth examining the coefficient. Hypothetically, if $\hat{\gamma}_{01}$ equaled the true γ_{01} then for every point increase in teacher evaluation score, there would be an almost seven point increase in student performance. Stated differently, for every standard deviation of increase in teacher evaluation score in math, the teacher's class's SAT-9 math performance would be predicted to increase only 2.92 points. Thus even if this point estimate was accurate, math teacher evaluation scores have a small relationship with students' math outcomes.

Another way to look at the significance of the mathematics teacher evaluation score is to examine the proportion of level 2 variance explained by the predictor. Calculations are: (118.41321-113.91646)/118.41321 = .0380, or less than 4% of the true between classroom variation in student achievement. Two possible statistical causes for the lack of effect of the teacher evaluation score in mathematics, as opposed to reading, are either that there is less variation in student achievement or

less variation in teacher evaluation scores in math than in reading. In fact, however, there was more variation in math than reading on both of these variables. This suggests that there is a difference in the predictive validity of the teacher evaluations in these subjects. The qualitative component of the larger study compared the evaluation system in these subjects, however, the findings are not reported here due to space constraints. The next section discusses the similar quantitative analysis in language arts.

Language Arts

The sample of students in language arts is smaller than that in reading and math because 53 students (mostly first graders) did not take the language test in 2000, and four students did not take the language section in 2001. Additionally, the concentration of missing data in first grade students in Spring 2000, made imputation unreliable. For this reason, the language arts results should be treated with a bit more caution than those in other subjects. Language arts was investigated in the same way as reading and math, beginning with the random intercept model:

Level 1: $ScaledLang01 = \beta_0 + \beta_1(Attendance)_{ij} + \beta_2(ImputeAttendance)_{ij} + \beta_3(Retain)_{ij} + \beta_4(SpecialEducation)_{ij} + \beta_5(ScaledReading00)_{ij} + \beta_6(ScaledMath00)_{ij} + \beta_7(ScaledLanguage00)_{ij} + \beta_8(ImputedReading)_{ij} + \beta_9(ImputedMath)_{ij} + \beta_{10}(ImputedELD)_{ij} + \beta_{11}(BeginningELD)_{ij} + \beta_{12}(EnglishProficient)_{ij} + \beta_{13}(ParentsNoHS)_{ij} + \beta_{14}(MissingParentEd)_{ij} + R_{ij}$

Level 2: $\beta_0 = \gamma_{00} + U_{0j}$

β_1 through β_{14} are fixed.

The results are presented in Table 9. In the level one analysis, the only interesting difference about this model compared to the others is that students whose parents did not graduate from high school scored an average of almost six points lower than other students, after controlling for all other factors.

Students with imputed math scores were left in the analysis because they were not statistically significantly different from the rest of the sample after controlling for other factors. The reliability of $\beta_0 = .828$ indicates that this model is quite reliable, even with the smaller sample size. This is potentially due to the fact that there are three pre-tests entered as control variables, which may give a more reliable estimate of students' academic proficiency in 2000. The chi-square

test for random effects of U_0 was significant, with a p-value of 0.000. This indicates that there is statistically different growth in achievement between classes. The intraclass correlation, $\rho = .24432$, showing that over 24% of the variation in student achievement is between classes. This is more than either of the other subjects.

Table 9: Results for the random intercept model for language arts

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
Intercept G00	158.7635	21.3576	7.434	0.000 ***
Attendance (grand centered) G10	0.0164	0.1015	0.162	0.872
Impute Attendance G20	-0.4935	9.5486	-0.052	0.959
Retain G30	-8.8240	3.8931	-2.267	0.023 *
Special Education G40	-20.4903	4.7413	-4.322	0.000 ***
Scaled Reading pre-test G50	0.3075	0.0459	6.704	0.000 ***
Scaled Math pre-test G60	0.1583	0.0408	3.878	0.000 ***
Scaled Language pre-test G70	0.3178	0.0538	5.905	0.000 ***
Impute Reading G80	4.4945	4.2579	1.056	0.292
Impute Math G90	-7.0604	9.7490	-0.724	0.469
Impute ELD G100	-0.9999	2.5921	-0.386	0.699
Beginning ELD G110	-6.4987	4.9407	-1.315	0.189
English Proficient G120	2.6140	3.1044	0.842	0.400
Parent No HS G130	-5.7941	2.4462	-2.369	0.018 *
Missing parent education G140	-9.6044	2.9749	-3.228	0.002 **
Reliability		**Variance Components**		
Reliability of $\beta_0 = .828$		τ_{00}= 134.0815, chi-square p-value 0.000		
Reporting *** = p < .001, ** = p < .01, * = p < .1		σ^2 = 414.7022		
Scale: a one point gain in reading on the Spring 2001 test. **Note**-in all tables G refers to γ		ρ = .24432		

How well do Vaughn's teacher evaluation scores predict the variation in student achievement? Table 10 below reports the Pearson correlations between the residuals and the average language, core five domains and overall scores.

Table 10: Correlation of Empirical Bayes' estimates of classroom effects in language arts with indicators of teacher quality

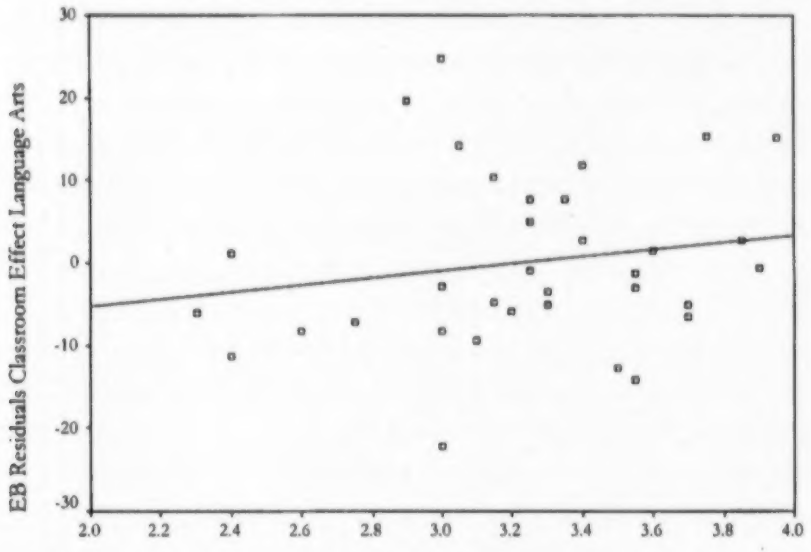
	Average Language	Average 5 core domains	Overall average
Pearson Correlation	.175	.194	.265

None of these are statistically significant predictors of classroom variation in language arts scores on the SAT-9. The poor correlation is also apparent when examining Figure 3 below, which shows the

BEST COPY AVAILABLE

correlation between Empirical Bayes' residuals of the intercept and teachers' average language arts evaluation score.

Figure 3: The correlation between Language TES and class performance in Language Arts



As the scatterplot shows, not only is the slope of the fit line relatively flat, but the spread of teacher evaluation scores around the line is larger than in any other subject, especially for teachers who are rated at least proficient (3.0). However, three outliers with received scores around 3.0 appear to make a strong contribution to lack of fit; without them, this correlation between the variables appears stronger.

After examining the correlations, the next step was to insert the average language teacher evaluation score into Level 2 of the Language model, as is represented by the equation:

Level 1: $(ScaledMath01) = \beta_0 + \beta_1(Attendance)_{ij} + \beta_2(ImputeAttendance)_{ij} + \beta_3(Retain)_{ij} + \beta_4(SpecialEducation)_{ij} + \beta_5(ScaledReading00)_{ij} + \beta_6(ScaledMath00)_{ij} + \beta_7(ImputedReading)_{ij} + \beta_8(ImputedELD)_{ij} + \beta_9(EarlyELD)_{ij} + \beta_{10}(EnglishProficient)_{ij} + \beta_{11}(ParentsNoHS)_{ij} + \beta_{12}(MissingParentEd)_{ij} + R_{ij}$

Level 2: $\beta_0 = \gamma_{00} + (AvgLanguageTES) + U_{0j}$

Table 11 reports the results.

Table 11: Language TES as a predictor of student language arts performance

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
Intercept G00	159.8920	21.3787	7.479	0.000 ***
Average Language TES (grand centered) G01	5.3397	5.2962	1.008	0.321
Attendance (grand centered) G10	0.0130	0.1016	0.128	0.898
Impute Attendance G20	-0.4902	9.5494	-0.051	0.959
Retain G30	-8.7468	3.8941	-2.246	0.025 *
Special Education G40	-20.4532	4.7418	-4.313	0.000 ***
Scaled Reading pre-test G50	0.3045	0.0460	6.626	0.000 ***
Scaled Math pre-test G60	0.1575	0.0408	3.855	0.000 ***
Scaled Language pre-test G70	0.3195	0.0538	5.933	0.000 ***
Impute Reading G80	4.7445	4.2644	1.113	0.266
Impute Math G90	-7.3762	9.7545	-0.756	0.450
Impute ELD G100	-0.9424	2.5926	-0.364	0.716
Beginning ELD G110	-6.3379	4.9426	-1.282	0.200
English Proficient G120	2.6570	3.1044	0.856	0.392
Parent No HS G130	-5.8683	2.4476	-2.398	0.017 *
Missing parent education G140	-9.5157	2.9757	-3.198	0.002 **
Reliability		**Variance Components**		
Reliability of β_0 = .827		τ_{00}= 133.43390		
Reporting *** = p < .001, ** = p < .01, * = p < .05		σ^2 = 414.79688		
Scale: a one point gain in reading on the Spring 2001 test. **Note**-in all tables G refers to γ		**Proportion of between class variation explained by language TES** = .0075		

As the results show, teachers' average language arts evaluation score was not a statistically significant predictor of student performance on the language arts portion of the SAT-9. This is confirmed by examining the proportion of between class variation in students' language achievement that was explained by average language teacher evaluation score, which was (134.0815-133.4339)/134.08152=.0075. This is especially interesting because, of the three subjects, language arts had the largest standard deviation for growth and the largest intraclass correlation both of which suggest that a good indicator of teacher quality would be likely to be a statistically significant predictor, all else being equal.

Discussion

In interpreting this data it is important to note two major limitations of the study. The sample was chosen to reflect a school that used an intensive teacher evaluation system embedded in a knowledge- and skills-based pay program. While Vaughn is an excellent case of knowledge- and

BEST COPY AVAILABLE

skills-based pay, the resulting sample is of marginal size for HLM. It is important, therefore, not to interpret non-significant results as meaning that there is no relationship between an indicator of teacher quality and student outcomes. It is also worth noting that as part of the broader study, both certification and years of experience were explored as level 2 predictors of variation in student achievement and neither of them were significant. In fact, small sample sizes are considered to be a problem in this area of study, which has often hindered researchers abilities to detect the effects of teacher characteristics (Greenwald et al., 1996; Wayne & Youngs, 2001).

Secondly, the current study is based on data from only two years of testing. Prior research has come to different conclusions about the consistency of classroom effects over time (Rowan, 2001; Heistad, 1999). The use of only two years of data creates a limitation of generalizability over time since it is unclear if classroom effects would be stable and if the evaluation system would have a similar relationship to the classroom effects over a longer time period. The study will be continued over the next few years to address this limitation.

Keeping the sample size and time limitations in mind, there is still a clear difference in the strength of the relationship between teacher evaluation scores and classroom achievement in reading compared with mathematics or language arts. The correlation between teacher evaluation scores and class achievement in reading of .545 is both significant and well above correlations between performance evaluation and outcomes found in earlier research (Heneman, 1986; Medley & Coker, 1987; Peterson, 2000). This establishes the predictive validity of the Vaughn teacher evaluation system in reading. Additional analyses (not reported here) of the relationship between a composite measure of student performance and the overall teacher evaluation score showed a strong and significant relationship for the overall measures as well, however, the results in mathematics and language arts were not as positive.

The implications of this study are that it is possible to improve teacher evaluation systems by specifying a clear definition of high quality teaching in specific subject areas and evaluating teachers based on that definition. The qualitative component of the study investigated properties of literacy

evaluation and instruction at Vaughn that led to higher consequential validity in reading than in other subjects. The lack of significant findings for the teacher evaluation system in math and language arts, like similar findings at Vaughn for teacher certification and years of experience, are inconclusive, which does not mean that there is no relationship. In fact, in the context of research on evaluation, the performance of the evaluation system in math and language arts is generally in line with correlations between evaluation and objective outcomes in other studies.

It is also important to consider several factors that may contribute to the lack of correlation, not all of which are inherently undesirable:

1. A large proportion of students at Vaughn are English Language Learners. This increases the measurement error in test scores, which may not accurately reflect student knowledge and skills. See, for example, August and Hakuta (1998) for more information on this. While the researcher acknowledges this potential problem, due to current California policy on assessing English Language Learners, this was unavoidable.
2. Data for this analysis were taken from two years of student testing. While some studies on the reliability of teacher effects across multiple years suggests that this may not be problematic (Heistad, 1999) it would be desirable to continue this study longitudinally due to score volatility.
3. As Porter and Smithson (2000) have shown, if the taught curriculum does not match the tested curriculum, student outcomes will not be a good reflection of teacher quality. It is possible that the reading curriculum at Vaughn is more aligned to the SAT-9 than the math or language arts curriculum.

This suggests that future research needs to compare teacher knowledge and skills and the teacher evaluation system at Vaughn in reading, math and language arts. The researcher took this direction in the second part of the study, which unfortunately cannot be reported here due to time and space constraints. Additionally, CPRE will continue to model the relationship between teacher evaluation scores and value-added measures of student outcomes at Vaughn over time. Finally, although the

results of this study are not fully generalizable beyond Vaughn, it suggests that more direct indicators of teacher skills (vs. more distal indicators like degrees and experience) have potential for finding the expected positive effects of teacher performance on student achievement. This study will be expanded in the future to include several large school districts to test the validity of teacher evaluation systems embedded in knowledge- and skills-based pay in a variety of sites and further explore the potential of innovative teacher evaluation systems as measures of teacher knowledge and skills.

Bibliography

August, D., & Hakuta, K. (1998). *Educating Language-Minority Children.* Washington, D.C.: Committee on Developing a Research Agenda on the Education of Limited-English-Proficient and Bilingual Students. Board on Children, Youth, and Families. Commission on Behavioral and Social Sciences and Education. National Research Council. Institute of Medicine.

Bond, L., Smith, T., Baker, W. K., & Hattie, J. A. (2000). *The Certification System of the National Board for Professional Teaching Standards: A Construct and Consequential Validity Study.* Greensboro: Center for Educational Research and Evaluation, The University of North Carolina at Greensboro.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How People Learn: Brain, Mind, Experience, and School.* Washington, D.C.: National Academy Press.

Brophy, J. (1986). Teacher Influences on Student Achievement. *American Psychologist*(October 1986), 1069-1077.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model. *American Journal of Education*(November), 65-108.

Cohen, M. (1983). Instructional, Management, and Social Conditions in Effective Schools. In A. Odden & L. D. Webb (Eds.), *School Finance and School Improvement Linkages for the 1980's* (pp. 17-50). Cambridge: Ballinger.

Coleman, J. S. (1990). *Equality and Achievement in Education.* Boulder: Westview Press.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Education Policy Analysis Archives, 8*(1), 50.

Darling-Hammond, L., & Ball, D. L. (1998). *Teaching for High Standards: What Policymakers Need to Know and Be Able to Do* (JRE-04): Consortium for Policy Research in Education. National Commission on Teaching and America's Future.

Dreeben, R. (2000). Structural Effects in Education: A History of an Idea. In M. T. Hallinan (Ed.), *Handbook of the Sociology of Education* (pp. 107-135). New York: Kluwer Academic.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The Effect of School Resources on Student Achievement. *Review of Educational Research, 66*(3), 361-396.

Hanushek, E. (1971). Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data. *The American Economic Review, 61*(2, Papers and Proceedings of the Eighty-Third Annual Meeting of the American Economic Association), 280-288.

Haycock, K. (1998). Good Teaching Matters: How Well-Qualified Teachers Can Close the Gap. *Thinking K-16, 3*(2), 1-14.

Heistad, D. (1999, April, 1999). *Teachers Who Beat the Odds: Value-Added Reading Instruction in Minneapolis 2nd Grade Classrooms.* Paper presented at the American Educational Research Association Conference, Montreal, CANADA.

BEST COPY AVAILABLE

Heneman, R. L. (1986). The Relationship between Supervisory Ratings and Results-oriented Measures of Performance: A Meta-analysis. *Personnel Psychology, 39*(4), 811-826.

Herman, J. L., Brown, R. S., & Baker, E. L. (2000). *Student Assessment and Student Achievement in the California Public School System. CSE Technical Report* (CSE-TR-519). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Heubert, J. P., & Hauser, R. M. (1999). *High Stakes: Testing for Tracking, Promotion and Graduation.* Washington, D.C.: National Research Council.

Kellor, E., Milanowski, T., Odden, A., & Gallagher, H. A. (2001). *How Vaughn Next Century Learning Center Developed a Knowledge- and Skill-Pay Program*: Consortium for Policy Research in Education. Wisconsin Center for Education Research, University of Wisconsin-Madison. http://www.wcer.wisc.edu/cprc/tcomp/research/ksbp/studies.asp

Little, R. J. A., & Rubin, D. R. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

Medley, D. M., & Coker, H. (1987). The Accuracy of Principals' Judgments of Teacher Performance. *Journal of Educational Research, 80*(4), 242-247.

Meyer, R. H. (1996). Value-Added Indicators of School Performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's Schools: The Role of Incentives* (pp. 197-123). Washington, D.C.: National Academy Press.

Meyer, R. H. (1997). Value-Added Indicators of School Performance: A Primer. *Economics of Education Review, 16*(3), 283-301.

Milanowski, A., & Gallagher, H. A. (2001). *Vaughn Next Century Learning Center Performance Pay Survey School Report*: Consortium for Policy Research in Education.

Nelson, B. S., & Sassi, A. (2000). Shifting Approaches to Supervision: The Case of Mathematics Supervision. *Educational Administration Quarterly, 36*(4), 553-584.

Odden, A., & Kelley, C. (2002). *Paying Teachers for What They Know and Do: New and Smarter Compensation Strategies to Improve Schools* (2nd ed.). Thousand Oaks: Corwin Press.

Peterson, K. D. (2000). *Teacher Evaluation: A Comprehensive Guide to New Directions and Practices* (Second Edition ed.). Thousand Oaks: Corwin Press.

Porter, A. C., & Brophy, J. (1988). Synthesis of Research on Good Teaching: Insights from the Work of the Institute for Research on Teaching. *Educational Leadership*(May 1988), 74-85.

Porter, A. C., & Smithson, J. L. (2000, April, 2000). *Alignment of State Testing Programs, NAEP and Reports of Teacher Practice in Mathematics and Science in Grades 4 and 8.* Paper presented at the American Educational Research Association, New Orleans, LA.

Porter, A. C., Youngs, P., & Odden, A. (2001). Advances in Teacher Assessments and Their Uses. In V. Richardson (Ed.), *Handbook of Research on Teaching* (4th ed., pp. 259-297). Washington, D.C.: American Educational Research Education.

Rogosa, D. (1999). *How Accurate Are the STAR National Percentile Rank Scores for Individual Students? An Interpretive Guide* [website]. National Center for Research on Evaluation, Standards, and Student Testing. Retrieved 2/29/00, 2000, from the World Wide Web: http://www.cse.ucla.edu

Rowan, B. (2001). *What Large-Scale, Survey Research Tells us About Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools.* Ann Arbor: Consortium for Policy Research in Education.

Rowan, B., Chiang, F.-S., & Miller, R. J. (1996). *Using Research on Employee Performance to Study Teaching Effectiveness: An Analysis of Teacher Effects on Student Achievement in Mathematics using NELS:88 Data.* Paper presented at the American Educational Research Association, New York.

Wayne, A. J., & Youngs, P. (2001, November 2, 2001). *Teacher Characteristics and Student Achievement Gains: A Review.* Paper presented at the Association for Public Policy Analysis and Management, Washington, DC.

Webster, W. J., Mendro, R. L., Orsak, T., Weerasinghe, D., & Bembry, K. (1997). Little Practical Difference and Pie in the Sky: A Response to Thum and Bryk and a Rejoinder to Sykes. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* (pp. 120-130). Thousand Oaks: Corwin Press.

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1998, April 13-17, 1998). *An Application of Hierarchical Linear Modeling to the Estimation of School and Teacher Effect.* Paper presented at the American Educational Research Association, San Diego, CA.

WestEd. (1998). *Case Study: Vaughn Next Century Learning Center.* Los Angeles: WestEd, Los Angeles Unified School District.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education, 11*(3), 57-67.

BEST COPY AVAILABLE

APPENDIX 1: DESCRIPTIVE STATISTICS

LEVEL-1 DESCRIPTIVE STATISTICS for literacy and math models

VARIABLE NAME	N	MEAN	SD	MINIMUM	MAXIMUM
FEMALE	584	0.51	0.50	0.00	1.00
ATTEN	584	174.23	8.82	61.00	180.00
IMPUATT	584	0.01	0.09	0.00	1.00
RETAIN	584	0.06	0.23	0.00	1.00
NS	584	0.04	0.19	0.00	1.00
SCLRD00	584	564.26	48.34	458.00	717.00
SCLMA00	584	574.61	43.47	460.00	737.00
SCLRD01	584	597.97	41.05	499.00	727.00
SCLMA01	584	610.74	40.65	506.00	762.00
IMPURD	584	0.07	0.26	0.00	1.00
IMPUELD	584	0.23	0.42	0.00	1.00
BELD	584	0.07	0.26	0.00	1.00
EPROF	584	0.21	0.41	0.00	1.00
NOHS	584	0.39	0.49	0.00	1.00
HSBEYOND	584	0.27	0.45	0.00	1.00
MISPARED	584	0.34	0.47	0.00	1.00

LEVEL-1 DESCRIPTIVE STATISTICS for language arts models

VARIABLE NAME	N	MEAN	SD	MINIMUM	MAXIMUM
ATTEN	532	174.15	9.09	61.00	180.00
IMPUATT	532	0.01	0.10	0.00	1.00
RETAIN	532	0.06	0.24	0.00	1.00
NS	532	0.04	0.20	0.00	1.00
SCLRD00	532	568.17	46.88	464.00	717.00
SCLMA00	532	578.54	41.67	471.00	737.00
SCLLA00	532	579.58	37.08	498.00	682.00
SCLRD01	532	600.97	39.88	499.00	727.00
SCLMA01	532	613.29	40.08	506.00	762.00
SCLLA01	532	601.99	37.91	517.00	712.00
IMPURD	532	0.07	0.25	0.00	1.00
IMPMATH	532	0.01	0.10	0.00	1.00
IMPUELD	532	0.23	0.42	0.00	1.00
BELD	532	0.06	0.23	0.00	1.00
EPROF	532	0.22	0.41	0.00	1.00
NOHS	532	0.37	0.48	0.00	1.00
MISPARED	532	0.40	0.49	0.00	1.00

LEVEL-2 DESCRIPTIVE STATISTICS

VARIABLE NAME	N	MEAN	SD	MINIMUM	MAXIMUM
YRSEXP	34	5.57	5.26	1.00	30.00
BCLAD	34	0.18	0.39	0.00	1.00
AVGELD	34	3.78	0.79	2.09	5.20
CRED	34	0.53	0.51	0.00	1.00
AVGLP	34	3.27	0.57	1.20	3.90
AVGCM	34	3.36	0.49	2.10	4.00
AVGLIT	34	3.24	0.39	2.30	3.85
AVGLANG	34	3.24	0.42	2.30	3.95
AVGMATH	34	3.13	0.42	2.25	4.00
AVG5	34	3.25	0.42	2.19	3.92

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Relationship between Measures of Teacher Quality and Student Achievement: The Case of Vaughn Elementary	
Author(s): H. Alix Gallagher	
Corporate Source: Consortium for Policy Research in Education University of Wisconsin-Madison	Publication Date: 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: H. Alix Gallagher	Printed Name/Position/Title: H. Alix Gallagher, Project Assistant
Organization/Address: 1025 W. Johnson St. 653G Madison, WI 53706	Telephone: (608)265-3523 FAX:
E-Mail Address: hagallagher@facstaff.wisc.edu	Date: 5/14/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)